Lexical Database of the Tibetan Grammatical Treatises Corpus*

Pavel Grokhovskiy, Maria Khokhlova, Maria Smirnova and Victor Zakharov

St. Petersburg State University, St. Petersburg, Russia

Abstract - The paper is devoted to Tibetan grammatical terminology. For this purpose Tibetan grammatical works corpus was created. At the same time Russian translations of the works were added to the corpus, so it is factually a parallel Tibetan-Russian corpus. The corpus represents the collection of grammar treatises of the Tibetan grammatical tradition formed in VII-VIII cc. The corpus is useful to researchers of the Tibetan linguistic tradition as well as to those specialized in linguistic studies of classical and modern Tibetan and its teaching. On the basis of corpus a specific grammatical lexical database is created. The database will be useful both to tibetologists and general linguistics specialists.

Index Terms - linguistic terminology, Tibetan language, corpus linguistics, parallel corpus, morphology, tagging, lexical database

1. Introduction

The project focuses on the creation of the Tibetan grammatical terminology database and Tibetan-Russian parallel corpus of traditional grammar treatises. The Tibetan linguistics is mainly based on grammars created by Buddhist scholars and thus is highly connected with Indian tradition. Methods of language description and analysis are greatly different from those of Western linguistics. Modern Tibetan scholars continue following and developing the Tibetan grammatical tradition. The corpus includes the basic grammar treatises and commentaries, which are considered to be the most important grammatical works within the Tibetan grammatical tradition. On the basis of the corpus a specific grammatical lexical database is created which will be useful both to tibetologists and general linguistics specialists.

2. Modern CORPUS Linguistics of the Tibetan Language

Despite the fact that scholars in different countries (Germany, Great Britain, People's Republic of China, USA and Japan) are engaged in working out of Tibetan texts corpus presentation, still there is no common standard for it. Last four conferences of the International Association for Tibetan Studies (IATS Seminar) included section «Tibetan Information Technology», where computer technology projects in the field of Tibetan Studies were represented. They also include projects focused on the creation of Tibetan corpus. The creation of Tibetan language corpora abroad has just begun. The cooperative research project 441 under the guidance of B. Zeissler in Eberhard Karls University (Tübingen, Germany) established the subproject B11 «Semantic roles, case relations, and cross-clausal reference in Tibetan» (2002-2008) [1]. In 2012 U. Pagel from Department of the Study of Religions and N. Hill from Department of China and Inner Asia and Departments of Linguistics began development of the Tibetan corpus which contained 1 million syllables. Its texts covers three historical periods of the Tibetan language: preclassical, classical and modern (more information see http://www.soas.ac.uk/news/newsitem73472.html As accessed on 29.03.2015).

The first difference between the corpus of the Tibetan traditional grammar treatises and the projects mentioned above is the development of special system of linguistic tags [2; 3] The second difference is related to the involved materials. All texts represent one of the traditional Tibetan sciences – linguistics.

3. Tibetan-Russian Parallel Corpus

The project has two main tasks: creation of parallel corpus of the Tibetan grammatical treatises with Russian translation and creation of a specific grammatical lexical database with frequency characteristics and semantic relations [4].

Tibetan texts and Russian translations in the corpus are aligned according to sentence boundaries of the Tibetan part. Word forms of the Tibetan part are also tagged. It should be noted that tokenisation of Tibetan texts is a sophisticated problem because according to the traditional Tibetan orthography only syllable borders are marked.

Tibetan text undergoes morphological tagging (lemmatization, part-of-speech tagging, grammatical annotation of verb forms, eliminating of grammatical homonymy). The Tibetan language tag system was developed. Corpus texts are also provided with metadata including information about genre, date of creation, author.

Every word is manually provided with the following data: word form in Tibetan script, word form in Latin transliteration, lemma (lexical item) in Tibetan script, lemma in Latin transliteration, part-of-speech tag (Table 2), terminological tag (Table 1). Further, TreeTagger, a program tool for automatic aligning, is supposed to be adapted to the Tibetan language. In this regard the corpus with manual aligning will be used as training corpus.

* The authors acknowledge Saint-Petersburg State University for a research grant 2.38.293.2014 Modernizing the Tibetan Literary Tradition for a study of the content of Tibetan grammar treatises. The model of linguistic data presentation in the parallel corpus and lexical database were developed with financial support of the Russian Foundation for Basic Research as a part of the research project 13–06–00621 "The Pilot Version of Tibetan Grammar Texts' Electronic Corpus".

TABLE I Fragment of aligned text

Word form (Tibe-tan script)	Word form (transli-teration)	Lemma (Tibetan script)	Lemma (transli-teration)	Part-of-speech tag	Termino-logical tag
<s>					
<align>					
དེ་	de	དེ་	de	P	
ནི་	ni	ནི་	ni	Top	
སྡུད་	sdud	སྡུད་	sdud	VN	Gram L TGrMark
དང་	dang	དང་	dang	Cj	
འབྱེད་པ་	'byed pa	ཕྱེད་	byed	VN	Gram L TGrMark
དང	dang	དང	dang	Cj	
།	//	།	//	Punct	
རྒྱུ་མཚན་	rgyu mtshan	རྒྱུ་མཚན་	rgyu mtshan	N	Gram GenLex TGrMark
ཚེ་སྐབས་	tshe skabs	ཚེ་སྐབས་	tshe skabs	N	Gram L TGrMark
གདམས་ངག་	gdams ngag	གདམས་ངག་	gdams ngag	N	Gram GenLex TGrMark
ལྔ	lnga	ལྔ	lnga	Num	
ཨོ	'o	ཨོ	'o	Fin	
།	//	།	//	Punct	
</align>					
</s>					

TABLE 2 Tags for Grammar Words

№	Tag	Grammar Word	Example
1	Cj	conjunction	dang
2	Pp	postposition	drung du
3	Erg	ergative marker	allomorphs kyis, gyis, gis, s, yis
4	Com	comitative marker	dang
5	Dat	dative marker	la
6	Loc	locative marker	na
7	Dest	destinative marker	allomorphs tu, du, ra, ru, su
8	Abl	ablative marker	las
9	El	elative marker	nas
10	Comp	comparative marker	allomorphs pas, bas
11	Gen	genitive marker	allomorphs kyi, gyi, gi, ‘i, yi
12	Fin	final particle	allomorphs go, ngo, do, no, bo, mo, ‘o, ro, lo, so, to
13	Top	topicalizing particle	ni
14	Ind	indefinite particle	allomorphs cig, zhig, shig
15	Emph	emphatic particle	allomorphs kyang, yang, 'ang
16	Quant	quantifiers	tsam, kho na, 'ba' zhig, snyed
17	Pl	plural marker	rnams
18	Quot	quotation marker	allomorphs ces, zhes

4. Lexical Database

A. Special Tagging of Grammatical Terminology

It is not typical for the Tibetan linguistics to emphasize such traditional subdisciplines of the Western linguistics as phonology, morphology and syntax. Basic terms of the Tibetan grammatical tradition denote basic units of different language levels [5].

Most Tibetan authors begin their grammatical works with the description of the Tibetan alphabet, different types of phonemes, rules of syllable composition and phoneme combination as well as rules of morphological combination of phonemes. Tibetan grammars also contains the description of functional words and morphemes.

The Tibetan linguistic tradition borrowed the Indian idea of seven cases. In Indian linguistics cases are connected with kāraka category which represents an intermediate level between semantics and morphology. This system of kāraka categories was also borrowed by Tibetans.

B. Tags for Grammatical Terminology

Elements of traditional grammatical metadescription such as terminological categories (phonological, morphological and syntax terms), Sanskrit equivalents for loans, links to synonyms, hyperonyms, hyponyms are added to the lexical database as well as scientific commentaries.

The use of special grammatical tags given in Table 3 makes it possible to divide different terminological fields: grammatical terminology (tag Gram) and terms of traditional sciences (tag GenScien).

Certain tags stand for models of terms’ formation: by terminologisation of common words (tag GenLex) or through borrowing (tag L).

The tag TBas is used for basic grammatical terminology. Polysemy is the main characteristic of the Tibetan terminology and basic grammatical terms in particular. Therefore one of the main tasks was to separate phonological terms (tag TPhon) and terms for different types of graphemes (tag TGra).

Grammatical terms imported from the corpus through the use of special grammatical tags form the lexical database of Tibetan grammatical terminology, which contains the additional information about origin language for loans, foreign equivalents, way of borrowing (phonetic or semantic borrowing, calquing, hybrid terms) etc.

TABLE 3 Tags for grammatical terminology

Characteristic of classification	Tag	Meaning
Terminological field	Gram	term of the Tibetan grammatical tradition
	GenScien	general scientific term
Origin	GenLex	term of Tibetan origin
	L	borrowed term
Type of terminology	TBas	basic grammatical term
	TPhon	phonological term
	TGra	grapheme type
	TGrMark	name of auxiliary morphemes and lexemes
	TCGr	case grammar term

C. Structure of Grammatical Lexical Database

Grammatical lexical database of the Tibetan grammatical treatises corpus contains lexical units selected from the Tibetan part of the corpus by appropriate tags.

TEI recommendations (Text Encoding Initiative) are taken as a methodological basis for database *exchange format* [6]. It is important that TEI has tags for relation links to create network data representation in linear XML files.

Let's describe grammatical lexical database representation template in XML format according to database structure. This template has several divisions and representation levels, namely, lexical unit level, link level and example level.

Lexical unit of the database in XML begins with record:

```
<entry n="1" type="lex">
<term>ཀླུ་ལི་</term>
<pron>A li</pron>
```

where index number of a lexical unit (n="1"), its type (type="lex" – lexical unit) and entry word ཀླུ་ལི་ in Tibetan script (tag <term>) and transcription (tag <pron>) are given.

It is followed by the block with grammatical information (tag <gramGrp>):

```
<gramGrp>
  <pos> N </pos>
</gramGrp>
```

which contains part-of-speech tag (tag <pos>) and additional grammatical information (tags <gen>, <flex> etc.).

The same level includes one or several etymological information blocks (tag <etym>):

```
<etym n="1">
  <lbl> Sumcupa </lbl>
  <date> 8th c. </date>
</etym>
```

which contains sequence number of the etymology block part-of-speech tag (tag <etym>), source of information (tag <lbl>), date of registration. Also tags <mentioned>, form of registration, and <lang>, language what a word was presumably borrowed, could be used.

Link level and example level are represented in the same way.

In the end data are converted from exchange format into Microsoft SQL on the Microsoft.NET platform.

D. User Interface

The database interface is a window application powered by Microsoft.NET and closely integrated with a system core. During the interface development the following tasks were set:

1) searching all lexemes in the database;
2) displaying lexemes in a convenient form;
3) manual adding of new lexemes;
4) editing of available lexemes;
5) loading (importing) lexical units records from XML to TEI format;
6) saving (exporting) database records into TEI format.

Functions of the grammatical lexical database are as follows:

1) statistical data provision where appropriate;
2) retrieval of an entry for a given word;
3) retrieval of synonyms for a given word;
4) retrieval of hyponyms for a given word;
5) retrieval of hyperonyms for a given word;
6) frequency of a given terminological tag;
7) retrieval of lexemes marked by a given terminological tag.

5. Conclusion and Future Works

Pilot version of the Tibetan grammatical treatises corpus will be useful to all researchers of Tibetan grammatical works. Nowadays there is no available Tibetan language corpora aligned and translated into Russian. Therefore the corpus also could be useful for linguistic researches, Tibetan language research and teaching.

A frequency dictionary of Tibetan lexical units (grammar terms) and semantic analysis of the lexical database will form a linguistic ontology that includes hyponyms and hyperonyms, polysemic words and synonyms.

All this will allow scholars to analyze the structure of terminological fields and estimate the degree to which common words were turned into technical terms.

In the future the corpus might be provided with syntactic annotation, extended and developed in a more extensive corpus of Tibetan texts including those on other traditional Tibetan sciences: Buddhist religious doctrine, logic, medicine, craft, poetics, synonymics, prosody, astrology and drama.

References

[1] Wagner, Andreas and Bettina Zeisler (2004). A syntactically annotated corpus of Tibetan. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisboa, May 2004.

[2] Grokhovskiy P.L. Kategorii skazuemosti i nominalizatsii deystviya (substantivno-ad'' ektivnye formy) v klassicheskom tibetskom yazyke // Ocherki po teoreticheskoy grammatike vostochnykh yazykov: Sushchestvitel'noe i glagol./Pod redaktsiey V.G. Guzeva. – SPb.: Izdatel'skiy dom SPbGU, 2001, S. 269–288.

[3] Grokhovskiy P.L. Grammatika imeni sushchestvitel'nogo v klassicheskom tibetskom yazyke // Ocherki po teoreticheskoy grammatike vostochnykh yazykov: Sushchestvitel'noe i glagol./Pod redaktsiey V.G. Guzeva. – SPb.: Izdatel'skiy dom SPbGU, 2001. C. 76–91.

[4] Grokhovskiy, P. L., Zakharov, V. P., Lebedeva, Yu. N., Smirnova, M. O., Khokhlova M. V.: "Korpus pamjatnikov tibetskoj grammaticheskoj traditsii [Corpus of the Tibetan traditional grammar treatises]". In: *Trudy mezhdunarodnoj konferentsii 'korpusnaja lingvistika-2013 [Proceedings of the InternatiomalConference "Corpus Linguistics-2013"]*, Sankt-Peterburg , Sankt-Peterburgskij gos. universitet, Filologicheskij fakultet, 2013, pp. 258–265.

[5] Smirnova M.O. "Bazovye terminy tibetskoy grammaticheskoy traditsii". In: *Vestnik Sankt-Peterburgskogo universiteta. Seriya 13. Vostokovedenie. Afrikanistika*. Vypusk 1, Sankt-Peterburg, Izdatel'stvo Sankt-Peterburgskogo gos. Universiteta, 2014, pp. 23–34.

[6] *TEI P5: Guidelines for Electronic Text Encoding and Interchange* / Eds. L. Burnard, S. Bauman. S. l. 2010.